$g_t(\mathbf{x}(t), \mathbf{u}(t), \mathbf{w}(t))$ A real-valued *reward function,* specifying the reward in period t as a function of the current state $\mathbf{x}(t)$, the decisions $\mathbf{u}(t)$, and disturbance $\mathbf{w}(t)$. The reward is assumed to be finite for all t. The total reward is additive,

$$g_{T+1}(\mathbf{x}(T+1)) + \sum_{t=1}^{T} g_t(\mathbf{x}(t), \mathbf{u}(t), \mathbf{w}(t)),$$

where $g_{T+1}(\mathbf{x}(T+1))$ is a *terminal reward.*

The objective is to maximize the total expected reward

$$E\left[g_{T+1}(\mathbf{x}(T+1)) + \sum_{t=1}^{T} g_t(\mathbf{x}(t), \mathbf{u}(t), \mathbf{w}(t))\right],$$

by choosing control actions **u**(1), **u**(2),..., **u**(T). We will assume that the functions $\mathbf{f}_t$, g_t and the disturbances $\mathbf{w}(t)$ are such that this expectation is always finite for any feasible sequence of control decisions.[1]

These control actions may be functions of the current state of the form $\mathbf{u}(t) = \mu_t(\mathbf{x}(t))$.[2] A collection of such functions $\{\mu_1, \mu_2, \ldots, \mu_T\}$ is called a *policy* and is denoted simply by μ. A policy is called *admissible* if $\mathbf{u}(t) \in U_t(\mathbf{x}(t))$ for all t and $\mathbf{x}(t) \in S_t$. The class of all admissible policies is denoted $\mathcal{M}$. For a given initial state $\mathbf{x}(1) = x$, the expected reward of a policy μ is

$$V_1^{\mu}(\mathbf{x}) = E\left[g_{T+1}(\mathbf{x}(T+1))) + \sum_{t=1}^{T} g_t(\mathbf{x}(t), \mu_t(\mathbf{x}(t)), \mathbf{w}(t))\right]. \qquad \text{(D.1)}$$

An optimal policy, denoted μ^*, is one for which

$$V_1^{\mu^*}(\mathbf{x}) = \sup_{\mu \in \mathcal{M}} V_1^{\mu}(\mathbf{x}).$$

The optimal expected reward is denoted simply $V_1(\mathbf{x})$, so

$$V_1(\mathbf{x}) = \sup_{\mu \in \mathcal{M}} V_1^{\mu}(\mathbf{x}).$$

The Principle of Optimality

The *principle of optimality,* due to Bellman [33], lies at the heart of dynamic programming. It is a strikingly simple idea; namely, that if a policy is optimal for the original problem stated above, then it must be optimal for any subproblem of this original problem as well. That is, define the reward-to-go for policy μ at time t by

$$V_t^{\mu}(\mathbf{x}) = E\left[g_{T+1}(\mathbf{x}(T+1))) + \sum_{s=t}^{T} g_s(\mathbf{x}(s), \mu_s(\mathbf{x}(s)), \mathbf{w}(s)) \,\middle|\, \mathbf{x}(t) = \mathbf{x}\right].$$

[1] For example, the expectation is always finite if the reward function, state space, and disturbance space are all bounded.

[2] Note that we have explicitly assumed here that it is sufficient that the control depend only on the current state $\mathbf{x}(t)$ and the current time t, and it does not need to depend on any other information about the *history* of the process up to time t. Such controls are called *Markovian controls.* Since the disturbances are independent over time and the system function only depends on the current state, disturbance, and control, one can show that there always exists an optimal Markovian policy, so it is sufficient to consider only policies of this form.